Analysis of the Our World In Data COVID-19 dataset report

Abstract

**OBJECTIVE** : Since the emergence of the covid-19 virus in December 2019, the world's epidemic situation has been very serious. This report takes Our World In Data COVID-19 dataset as the research object. Aims to explore the current covid-19 situation in all countries in the world, the covid-19 trend in the four worst European countries, and the prediction of covid-19 mortality rate.

**METHODS** : The author uses a series of techniques to achieve the goal. Including spatial data visualization, Multi-line chart, multiple linear regression. The Our World In Data COVID-19 dataset and the UK COVID-19 Data dataset are used.

**RESULTS**: This study shows that the European region currently has the highest number of cases per million of any region in the world. The top four countries with the most severe outbreaks in Europe are the UK, Russia, France and Germany which are also on the rise of total cases. In addition, the mortality rate can be predicted by the number of total cases per million, the positive rate, the number of first doses of vaccination and the number of fully vaccinated people.

**CONCLUSION**:The research contributes to understanding the impact and trends of covid-19 two years after its development, and to help understand the factors that affect covid-19 mortality rate. But the main limitation is the time restriction, more time can help the author to improve the regression model and try other models to analyze the covid-19 data more effectively to help predict mortality rate.

Zhibo Liu

Reg number: 210164972

Word Count: 2746

**Content**

[1. Introduction and aim 4](#_Toc26322)

[2. Methodology 5](#_Toc31748)

[2.1 Visualisation of world epidemic situation on map 5](#_Toc7647)

[2.1.1 Gathering data 5](#_Toc11499)

[2.1.2 Structuring data 5](#_Toc5924)

[2.1.3 Visualising data 5](#_Toc29393)

[2.2 Trend of Covid-19 data 6](#_Toc12146)

[2.2.1 Gathering data 6](#_Toc28293)

[2.2.2 Structuring data 6](#_Toc26262)

[2.2.3 Visualising data 6](#_Toc25078)

[2.3 Prediction of covid-19 mortality rate 7](#_Toc31440)

[2.3.1 Gathering data 7](#_Toc11840)

[2.3.2 Structuring data 7](#_Toc32036)

[2.3.3 Exploring data 8](#_Toc27471)

[2.4 Summary 8](#_Toc3319)

[3. Results and discussion 9](#_Toc11633)

[3.1 The most severe areas of COVID-19 in the world 9](#_Toc7643)

[3.2 Trend of covid-19 in 4 European countries 10](#_Toc28807)

[3.3 The multiple linear regression model of covid-19 mortality rate 11](#_Toc119)

[4. Conclusion 13](#_Toc7609)

[References 14](#_Toc28652)

[R code 15](#_Toc2012)

# Introduction and aim

Beginning with the first emergence of the covid19 virus in Wuhan, China in December 2019, the World Health Organization (WHO) declared the virus a pandemic in March 2020 (Sanyaolu et al., 2020; Green, 2020). With its sudden appearance and constant human-to-human transmission, the number of infections has been growing rapidly. covid-19's unpredictable and massive worldwide spread poses a challenge for various countries in terms of blockades and medical systems (Sanyaolu et al., 2020; Sheng et al., 2021; Wolkewitz & Puljak, 2020). In response, countries around the world are making enormous efforts to combat the disease. However, the virus is not developing in the direction one would hope. A new coronavirus variant, Delta (Wise, 2021), was discovered in India in May 2021, followed by a new coronavirus variant, Omicron (Karim & Karim, 2021; Torjesen, 2021), which is more transmissible and has a higher viral load, in November 2021. Thus, the world epidemiological situation has accelerated even more. Therefore, an understanding and analysis of the covid-19 situation are necessary. Latif et al. (2020) indicate to the use of various methods and tools of data science, including statistical modelling data visualisation, can be used to extract insights from the acquired data. The first goal of this article is therefore to visualise the latest world covid-19 data on a world map to understand more about the changing to new coronaviruses. After visualisation of the data, it was possible to visually identify the areas of the world with the most severe covid-19 situation. The second objective of this article is to select the four countries with the highest number of covid-19 cases in the most severe regions of the world and analyse the trends in their outbreaks over the past year.

How the severity of covid-19 is measured is mainly by accidental infection and mortality (Villani et al., 2020). It is common to measure the severity of the virus through mortality rates. Therefore, there is a need to study mortality in the context of the changing impact of the covid-19 virus. The third objective of this article is therefore to predict covid-19 mortality in the UK. This article will cover in turn introduction and aim, methodology, results and discussion, and conclusion.

# **Methodology**

This section presents the methodology used to accomplish the three objectives of this article. Each sub-section consists of three elements, which are gathering data, structuring data, exploring data or visualising data. In this article, the author used Our World in Data COVID-19 Dataset (https://ourworldindata.org/coronavirus) to perform the analysis, which includes data on COVID-19 for countries around the world from February 2020 to December 12, 2021, such as total cases. All analyses were done using R studio, R version 4.1.1 (2021-08-10).

## **Visualisation of world epidemic situation on a map**

This section aims to find the latest COVID-19 situation for each country in the dataset, for example, the total number of cases per million people, and to visualise the data on a world map.

### **Gathering data**

For the epidemic data, the authors first read the dataset "owid-covid-data.csv" using the read.csv() function and named it owid\_covid\_data. To visualise the data on a map, it is necessary to obtain data from a world map. This can be done with the map\_data function included in the ggplot2 package. So after loading the ggplot2 package, use map\_data() to get the world map data and name it world\_map.

### Structuring data

Load the tidyverse package and, as the original data contains many unnecessary variables, use the select() function to select the "location", "date", "total \_cases\_per\_milliion" variables. In order to get the most recent data, use the filter() function to filter all data with "date" equal to "2021-12-12", recorded as world\_data\_20211212. However, the name of the USA in this dataset does not match the name in the world\_map dataset, which will cause problems, so it needs to be changed to "USA".

In order to visualise the data, the world\_data\_20211212 data and the world\_map data have to be merged. Therefore, here the author uses left\_join(world\_map,world\_data\_20211212,by = c("region"="location")) to merge the world\_map dataset based on the "region" of the world\_map dataset and saved it as world\_coviddata data.

### Visualising data

Loading the ggplot2 and maps packages, first use ggplot(world\_coviddata,aes(long,lat,group=group)).

Then use geom\_polygon() to draw the shape and fill in the graph with total cases per million. Moreover, use the scale\_fill\_viridis\_c() function to give the map colour by value and finally add the title. Therefore, this map shows the current number of COVID-19 cases in the world and it is easy to identify the areas with higher total cases per million.

## 2.2 Trend of Covid-19 data

Based on the analysis in the previous section, the purpose of this section is to find the four countries with the highest number of total COVID-19 cases in Europe by December 12, 2021, and then to investigate their trends over the past year.

### 2.2.1 Gathering data

Use the data owid\_covid\_data from the previous section.

### 2.2.2 Structuring data

Loading the tidyverse package. Use the select() function to select "continent", "location", "date", " total\_cases" variables from owid\_covid\_data data and use the filter() function to select the data with a continent equal to "Europe" and a date equal to "2021-12-12 " and record the data as europe\_country\_total\_cases\_latest data. The author then uses the arrange() function to sort the data by the "total\_cases" variable from largest to smallest and then uses the head() function to find the top four countries:

> head(europe\_country\_total\_cases\_latest)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | continent | location | date | total\_cases |
| 1 | Europe | United Kingdom | 2021-12-12 | 10881189 |
| 2 | Europe | Russia | 2021-12-12 | 9842021 |
| 3 | Europe | France | 2021-12-12 | 8362843 |
| 4 | Europe | Germany | 2021-12-12 | 6546010 |

Therefore, the data for these four countries are filtered from europe\_country\_total\_cases\_latest data using filter(). After changing the class of the date variable in the data using as.Date(), the data from 12 December 2020 to 12 December 2021 is filtered and saved as fourcountry\_data.

### 2.2.3 Visualising data

Loading ggplot2 package. Use ggplot() and geom\_line() function to build the line and Use aes(colour=location) in geom\_line() to assign different colours to the line graph by location:

ggplot(fourcountry\_data,aes(date,total\_cases))+

geom\_line(size=1,aes(colour=location))

Then, the author used the scale\_x\_date() function to set the label of the date to year-month-day format and also set the time interval to two months. Finally changing the axes labels and the image title. Accordingly, this line chart can clearly compare and present the trends of covid-19 epidemic data in four countries over the past year.

## 2.3 Prediction of covid-19 mortality rate

This section achieves the prediction of mortality rate from COVID-19 data by creating multiple linear regressions with multiple variables. Some of the variables are derived from the previous part of the data owid\_covid\_data. Another part of the variables is derived from the new data set: "UK COVID-19 Data". This dataset is from Kaggle.com and includes the total number of people who received the first dose of vaccine and the total number of people who received the second dose of vaccine from the start of 2021 to 12 December 2021.

### 2.3.1 Gathering data

Loading the tidyverse package. Use the select() function to select the required variables in the owid\_covid\_data data and filter the UK data from the start of 2021 to 12 December 2021 using filter(). Then use the mutate() function to add the new variable "mortality\_rate" and the mortality rate is calculated by dividing the total number of deaths per million by the total number of cases per million:

mutate(uk\_coviddata,mortality\_rate = total\_deaths\_per\_million / total\_cases\_per\_million)

Then, read the "UK\_National\_Total\_COVID\_Dataset.csv" data using read.csv(). Use select() to select the required part and save it as vaccinated\_data.

### 2.3.2 Structuring data

Loading the reshape package. To make variable names more concise, use the rename() function to change the variable name, e.g:

vaccinated\_data<-rename(vaccinated\_data,c(cumPeopleVaccinatedFirstDoseByPublishDate="total\_people\_vaccinated\_1stDose"))

Use the left\_join() method to merge the uk\_coviddata dataset with the vaccinated\_data dataset, saving it as the uk\_coviddata\_regression dataset. Analysis of the dataset using summary() reveals that there are missing values. So first use the anyNA() function to detect if there are missing values, then use the is.na() function to see where the missing values are, and finally, use the na.omit() function to remove the missing values. The author loads the GGally package and uses the ggpairs() function in the package to check the correlation of the variables of the uk\_ coviddata\_regression data. Through correlation detection, some uncorrelated variables can be removed by understanding the correlation between two variables before modelling.

### 2.3.3 Exploring data

After structuring the data, the authors deleted some variables, such as population density, and used the lm() function to build a model between mortality rate and total cases, new cases, total deaths, positive rate, the total number of people who are fully vaccinated, total people who vaccinated the first dose, and total people who vaccinated the second dose. In addition, another stepwise regression model was built using the backward option in the step() function to remove insignificant variables to improve the model. Finally, the parameters generated by the summary() function, such as R squared and p-value, are used to analyze the quality of the model and browse the results of the model.

## 2.4 Summary

Table 1 summarises the techniques used, the data set used, the variables used and the R package used in each section.

|  |  |  |  |
| --- | --- | --- | --- |
| **Technique** | **Dataset** | **Variable** | **R Package** |
| Spatial data visualisation | Our World In Data  World\_map | total\_cases\_per\_million | “tidyverse”  “ggplot2”  “maps” |
| Multi-line chart | Our World In Data | location  date  total\_cases | “tidyverse”  “ggplot2” |
| Multiple linear  regression | Our World In Data  UK COVID-19 Data | location  date  mortality\_rate  total\_cases\_per\_million  new\_cases\_smoothed\_per\_million  total\_deaths\_per\_million  positive\_rate  people\_fully\_vaccinated  total\_people\_vaccinated\_1stDose  total\_people\_vaccinated\_2ndDose | “tidyverse”  “ggplot2”  “reshape”  “GGally” |

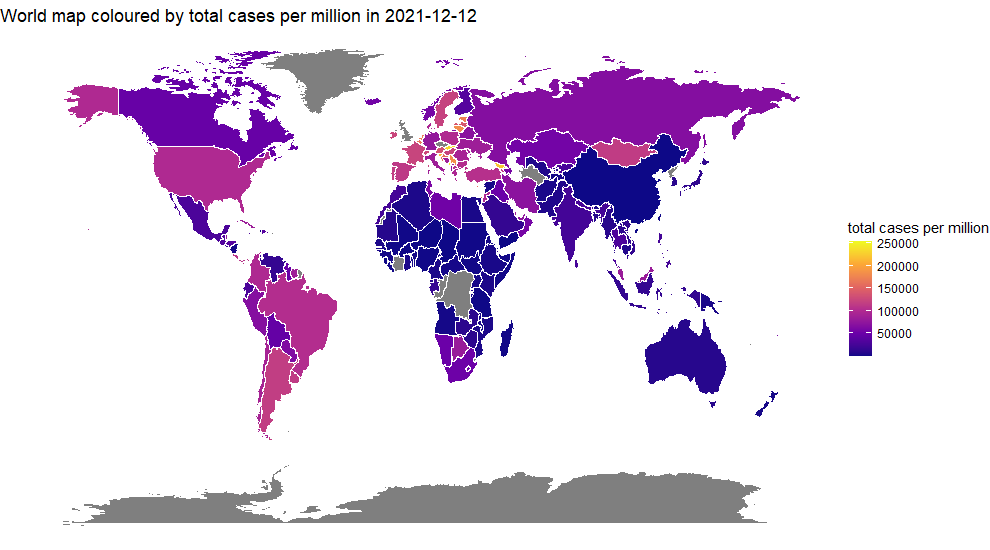
Table 1 Methodology

# Results and discussion

## The most severe areas of COVID-19 in the world

After obtaining the data using the operations described in the methodology, the authors produced the world map in Figure 1 which is coloured by total cases per million covid-19 on 12 December 2021. Figure 1 clearly shows the differences in the total cases per million available in different regions of the world. It is clear from the map that Europe is generally "brighter" than other regions or continents of the world, which is another way of saying that there are more total COVID-19 cases per million. This may be due to two reasons, the first one is that European countries do not have a higher total population and therefore have a higher number of total cases per million compared to the United States, which has a similarly high population and the number of total COVID-19 cases. The second is that, in line with Villani et al. (2020), European countries were inexperienced in facing the covid-19 virus during the initial phase of the pandemic and were, therefore, the most affected continent.

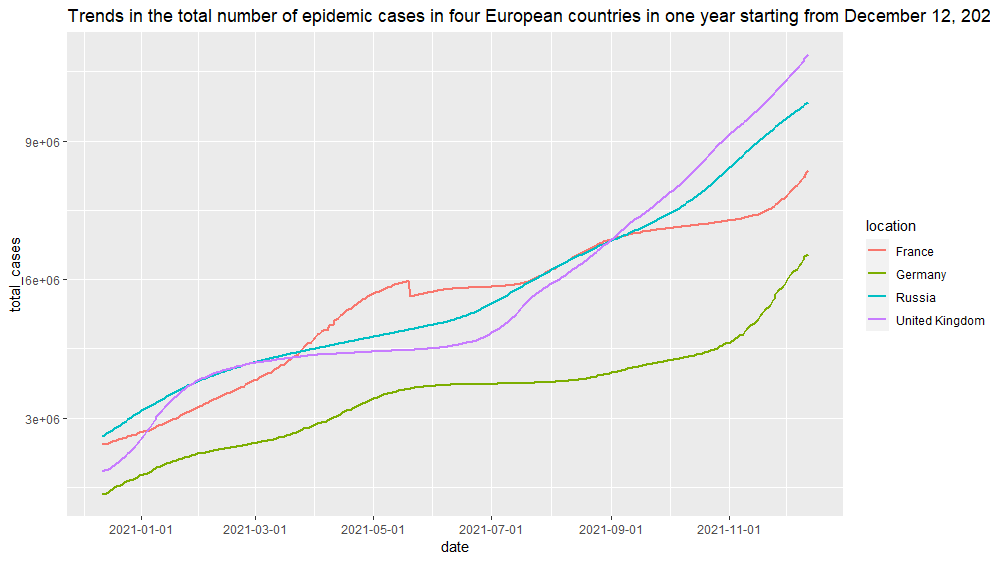
Figure 1



## Trend of covid-19 in 4 European countries

After the analysis of the world epidemic situation in the previous section, the authors found that the situation of covid-19 is more serious in the European region. Therefore, in this section, the top four European countries in terms of the total number of cases were selected in order of the total number of cases so far, which are the United Kingdom, France, Germany and Russia. Figure 2 shows the change in the number of cases from the four countries starting from December 12, 2020, to December 12, 2021. It can be seen that there is an overall increasing trend in the number of cases in all four countries. However, the number of cases in the four countries remained almost unchanged during the two months from May 1, 2021, to July 1, 2021. However, as indicated by Wise (2021) to the emergence of the new coronavirus variant Delta, the influx into Europe leads to a significant and rapid increase in the number of cases in the UK and Russia starting in July 2021. The subsequent emergence of the new coronavirus variant Omicron in November (Karim & Karim, 2021; Torjesen, 2021) continued to lead to an increase in the number of cases until December 12, when the UK became the country with the highest number of covid-19 cases in Europe.

Figure 2



## The multiple linear regression model of covid-19 mortality rate

In the previous section of the analysis, it was learned that among European countries the UK has seen a more aggressive trend in case of development over the past year and already has the largest number of COVID-19 cases in Europe in December 2021. In addition, Villani et al. (2020) indicated that people prefer to use mortality rates to measure the severity of an outbreak and that early knowledge of mortality rates can help to develop strategies. Therefore, this section aims to model mortality to aid prediction. So after processing the data a first model was built using multiple linear regression and the results of model 1 are displayed in Table 2.

Table 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| (Intercept) | 3.644e-02 | 2.258e-03 | 16.136 | < 2e-16\*\*\* |
| total\_cases\_per\_million | -1.295e-07 | 6.702e-09 | -19.325 | < 2e-16 \*\*\* |
| new\_cases\_smoothed\_per\_million | -1.051e-06 | 7.748e-07 | -1.356 | 0.1759 |
| total\_deaths\_per\_million | 5.209e-07 | 1.586e-06 | 0.328 | 0.7428 |
| positive\_rate | -4.409e-02 | 1.096e-02 | -4.021 | 7.20e-05 \*\*\* |
| people\_fully\_vaccinated | -8.797e-11 | 1.208e-11 | -7.280 | 2.52e-12 \*\*\* |
| total\_people\_vaccinated\_1stDose | 3.639e-11 | 1.914e-11 | 1.902 | 0.0581 . |
| total\_people\_vaccinated\_2ndDose | NA | NA | NA | NA |

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0005952 on 326 degrees of freedom

Multiple R-squared: 0.9885, Adjusted R-squared: 0.9883

F-statistic: 4678 on 6 and 326 DF, p-value: < 2.2e-16

Table 2 shows that the F-statistic p-value is less than 2.2e-16, which means that the p-value is less than 0.05, and the R-squared is 0.9883. This indicates that the quality of model 1 is good and that regression model 1 explains 98% of the data. However, there are some of these variables which have a coefficient that is greater than 0.05, such as the new cases smoothed per million and the total deaths per million variables. This may have an impact on the accuracy of the model. Therefore, the author chose to use the stepwise regression algorithm to improve model 1 with the backward option. Variables with high p values were removed using the backward method. Thus, the improved multiple regression model model 2 is shown in table 3.

Table 3

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| (Intercept) | 3.741e-02 | 3.130e-04 | 119.511 | < 2e-16\*\*\* |
| total\_cases\_per\_million | -1.287e-07 | 2.086e-09 | -61.698 | < 2e-16\*\*\* |
| positive\_rate | -5.638e-02 | 3.454e-03 | -16.325 | < 2e-16\*\*\* |
| people\_fully\_vaccinated | -8.743e-11 | 1.071e-11 | -8.164 | 7.02e-15\*\*\* |
| total\_people\_vaccinated\_1stDose | 3.434e-11 | 1.351e-11 | 2.542 | 0.0115\* |

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.000595 on 328 degrees of freedom

Multiple R-squared: 0.9885, Adjusted R-squared: 0.9883

1. statistic: 7020 on 4 and 328 DF, p-value: < 2.2e-16

From Table 3, it can be seen that the improved model 2 has reduced some variables compared to model 1. Three variables, smoothed new cases per million, total deaths per million, and the total number of people who received a second dose of vaccine, were removed because of their low impact on the model. When looking at the summary of model 2, it is possible to find that the R-squared is still 0.9883 and the p-value of the F-statistic is also less than 0.05. Therefore, the quality of model 2 is also very good. And the coefficients of all variables are less than 0.05, so combining the analysis of all the summary parameters, the author can reject the original hypothesis that the number of total cases per million, the positive rate, the number of people vaccinated with the first dose and the number of people fully vaccinated do not affect mortality rate. The results showed that the number of total cases per million, the positive rate, the number of first doses of vaccine and the number of fully vaccinated persons were significant, and that model 2 can explain 98.83% of the data. To summarize, covid 19 mortality rate was significantly influenced by the number of total cases per million, the positive rate, the number of first doses of vaccine and the number of fully vaccinated persons, and an increase in the number of persons receiving only the first dose of coronavirus vaccine contributed to the increase in mortality. The regression model is: Mortality rate = 3.741\*10-2 - 1.287\*10-7\* total\_cases\_per\_million - 5.638\*10-2 \* positive\_rate - 8.743\*10-11\*people\_fully\_vaccinated + 3.434\*10-11\* total\_people\_vaccinated\_1stDose.

# Conclusion

By taking covid-19 epidemic data as the research target, this study uses the new coronavirus dataset from Our World In Data dataset. The three objectives of this study were to map the world of covid-19, analyze the trends of covid-19 virus in the past year in four European countries and predict the covid-19 mortality rate in the UK. This study shows that the European region currently has the highest number of cases per million of any region in the world. The top four countries with the most severe outbreaks in Europe are the UK, Russia, France and Germany. The number of cases in all four countries is on the rise from December 2020 to December 2021, especially when the emergence of two new coronavirus variants, Delta and Omicron, has led to a significant increase in the number of cases in countries such as the United Kingdom. In addition, the mortality rate can be predicted by the number of total cases per million, the positive rate, the number of first doses of vaccination and the number of fully vaccinated people. The main limitation of this study is the time constraint, if more time was available, the validity of the model could be improved by improving the use of regression models and comparing different modelling approaches to analyze covid-19 data more efficiently.

# References

Green. (2020). Emerging diseases, re‐emerging histories. Centaurus, 62(2), 234–247. <https://doi.org/10.1111/1600-0498.12306>

Karim, & Karim, Q. A. (2021). Omicron SARS-CoV-2 variant: a new chapter in the COVID-19 pandemic. The Lancet (British Edition), 398(10317), 2126–2128. <https://doi.org/10.1016/S0140-6736(21)02758-6>

Latif, Usman, M., Manzoor, S., Iqbal, W., Qadir, J., Tyson, G., Castro, I., Razi, A., Boulos, M. N. K., Weller, A., & Crowcroft, J. (2020). Leveraging Data Science to Combat COVID-19: A Comprehensive Review. IEEE Transactions on Artificial Intelligence, 1(1), 85–103. https://doi.org/10.1109/TAI.2020.3020521

Sanyaolu, Okorie, C., Hosein, Z., Patidar, R., Desai, P., Prakash, S., Jaferi, U., Mangat, J., & Marinkovic, A. (2021). Global Pandemicity of COVID-19: Situation Report as of June 9, 2020. Infectious Diseases, 14, 1178633721991260–1178633721991260. https://doi.org/10.1177/1178633721991260

Sheng, Amankwah‐Amoah, J., Khan, Z., & Wang, X. (2021). COVID‐19 Pandemic in the New Era of Big Data Analytics: Methodological Innovations and Future Research Directions. British Journal of Management, 32(4), 1164–1183. <https://doi.org/10.1111/1467-8551.12441>

Torjesen. (2021). Covid-19: Omicron may be more transmissible than other variants and partly resistant to existing vaccines, scientists fear. BMJ (Online), 375, n2943–n2943. <https://doi.org/10.1136/bmj.n2943>

Villani, McKee, M., Cascini, F., Ricciardi, W., & Boccia, S. (2020). Comparison of deaths rates for COVID-19 across Europe during the first wave of the COVID-19 pandemic. Frontiers in Public Health, 8, 620416–620416. https://doi.org/10.3389/fpubh.2020.620416

Wise. (2021). Covid-19: Delta variant doubles risk of hospital admission compared with alpha variant, study shows. BMJ (Online), 374, n2152–n2152. https://doi.org/10.1136/bmj.n2152

Wolkewitz, & Puljak, L. (2020). Methodological challenges of analysing COVID-19 data during the pandemic. BMC Medical Research Methodology, 20(1), 81–81. https://doi.org/10.1186/s12874-020-00972-6

# R code

#\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*#

# Visualisation of world epidemic situation on map

#\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*#

1. Gathering data

library(maps)

owid\_covid\_data <- read.csv("owid-covid-data.csv", header = TRUE)

world\_map<-map\_data("world")

1. Structuring data

library(tidyverse)

world\_covid\_data <- owid\_covid\_data %>% select(location,date,total\_cases\_per\_million)

world\_data\_20211212 <- filter(world\_covid\_data,date == "2021-12-12")

View(world\_data\_20211212)

world\_data\_20211212[136,1] = "USA" # make the variable names in the two data the same

world\_coviddata <- left\_join(world\_map,world\_data\_20211212,by = c("region"="location"))# join two data based on region equal to location

View(world\_coviddata)

1. Visualising data

library(ggplot2)

ggplot(world\_coviddata,aes(long,lat,group=group))+

geom\_polygon(aes(fill=total\_cases\_per\_million),colour="white")+

scale\_fill\_viridis\_c(option = "plasma")+

theme\_void()+

labs(fill="total cases per million",title = "World map coloured by total cases per million in 2021-12-12")

#\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*#

# Trend of Covid-19 data

#\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*#

1. Structuring data

library(tidyverse)

library(ggplot2)

country\_total\_cases <- owid\_covid\_data %>% select(continent,location,date,total\_cases)

europe\_country\_total\_cases\_latest<-filter(country\_total\_cases,

continent == "Europe" & date =="2021-12-12")

europe\_country\_total\_cases\_latest<- arrange(europe\_country\_total\_cases\_latest,

desc(total\_cases))# sort by the number of total cases from largest to smallest

head(europe\_country\_total\_cases\_latest)

fourcountry\_data<-filter(europe\_country\_total\_cases,

location == "United Kingdom"| location=="Russia"|location=="France"|location=="Germany" )

class(fourcountry\_data$date)

fourcountry\_data$date<-as.Date(fourcountry\_data$date) # convert data format

fourcountry\_data<-filter(fourcountry\_data,

date>="2020-12-12"& date<="2021-12-12")

1. Visualising data

ggplot(fourcountry\_data,aes(date,total\_cases))+

geom\_line(size=1,aes(colour=location))+

scale\_x\_date(date\_labels = "%Y-%m-%d",date\_breaks = "2 month")+ # Change the format of the time label and set it to an interval of two months

labs("date","total cases",title = "Trends in the total number of epidemic cases in four European countries in one year starting from December 12, 2020")

#\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*#

# Prediction of covid-19 mortality rate

#\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*#

1. Gathering data

library(tidyverse)

library(ggplot2)

uk\_coviddata <- owid\_covid\_data %>% select(location,date,

total\_cases\_per\_million,

new\_cases\_smoothed\_per\_million,

total\_deaths\_per\_million,

positive\_rate,

people\_fully\_vaccinated,

population\_density)

uk\_coviddata<-filter(uk\_coviddata,location=="United Kingdom")

uk\_coviddata$date<-as.Date(uk\_coviddata$date)

uk\_coviddata<-filter(uk\_coviddata,date>="2021-01-01"&date<="2021-12-12")

uk\_coviddata<-mutate(uk\_coviddata,mortality\_rate = total\_deaths\_per\_million / total\_cases\_per\_million) # add new variable mortality\_rate

UK\_national\_covid\_19 <-read.csv("UK\_National\_Total\_COVID\_Dataset.csv",header=TRUE)

vaccinated\_data<-UK\_national\_covid\_19%>%select(date,areaName,

cumPeopleVaccinatedFirstDoseByPublishDate, cumPeopleVaccinatedSecondDoseByPublishDate)

1. Structuring data

library(reshape)

vaccinated\_data<-rename(vaccinated\_data,c(cumPeopleVaccinatedFirstDoseByPublishDate="total\_people\_vaccinated\_1stDose")) # change name of variable

vaccinated\_data<-rename(vaccinated\_data,c(cumPeopleVaccinatedSecondDoseByPublishDate="total\_people\_vaccinated\_2ndDose"))

vaccinated\_data$date<-as.Date(vaccinated\_data$date)

uk\_coviddata\_regression<-left\_join(uk\_coviddata,vaccinated\_data,by=c("date"="date"))

summary(uk\_coviddata\_regression)

anyNA(uk\_coviddata\_regression) # check if there are missing values

is.na(uk\_coviddata\_regression) # find where are missing values

uk\_coviddata\_regression<-na.omit(uk\_coviddata\_regression) # remove missing values

summary(uk\_coviddata\_regression)

library(GGally)

ggpairs(uk\_coviddata\_regression) # Check the correlation between variables

1. Exploring data

model1=lm(mortality\_rate~ total\_cases\_per\_million+new\_cases\_smoothed\_per\_million+total\_deaths\_per\_million+positive\_rate+people\_fully\_vaccinated+total\_people\_vaccinated\_1stDose+total\_people\_vaccinated\_2ndDose,data = uk\_coviddata\_regression)

summary(model1)

model2<-step(model1,direction = "backward") # use the stepwise regression algorithm to improve model 1 with the backward option

summary(model2)